MACHINE LEARNING FOR HEALTHCARE - FINAL PROJECT

Noa Mark & Yael Einy
Department of Computer Science
Tel Aviv University
{noamark, yaeleiny}@mail.tau.ac.il

1 INTRODUCTION

ICU targets prediction is important to achieve smart healthcare management. In this work,[1] we focus on predicting 3 different ICU targets: Mortality-during hospitalization or up to 30 days after discharge, Prolonged stay: length of stay $>$ 7 days and Hospital readmission- in 30 days after discharge, using MIMIC-III data set Alistair E.W. Johnson (2016).

2 COHORT DESCRIPTION

The initial cohort size stand on 40156 samples (including more than first admission), but due to our inclusion and exclusion criteria we train on 15270 samples.

All our targets suffer from imbalance distribution, where the positive class is much smaller than the negative. Prolong stay true class stand on around 32%, mortality 12% and hospital readmission 4%. Due to this fact, hospital readmission prediction can be seen as predicting tail events. We will attend this issues in method section.

In this work we used Johnson et al. (2017) and Shirly Wang (2020) important works for accessibility of medical concepts and prepossessing of the MIMIC-III data set. These repository ease the processes of cleaning edge cases and combining features into medical concepts. By that, allowing the researcher to focus on meaningful new contribution for the scientific community.

From Johnson et al. (2017) we extract all labs and vitals hourly mean and patient statics data. Note, they cover large scope of labs and vitals. They also provide another modality - intervention - 14 categorical variable in hourly resolution that describe different medical interventions. (see Appendix B). The static data on the patient include type of admission and type of insurance that we will also use for our prediction (see data exploration).

From Shirly Wang (2020) we extract different severity scores calculating based on the 24 first hours of admission data: SIRS, SAPS-II, APS-III, SOFA, LOADS and OASIS. We also extract MELD score, a score often used to assess health of liver transplant candidates and related to organ failure, computed based on the first day of admission data. Its important to note, medicine are using MELD to assess 3-month mortality chance due to liver disease. Therefore, it can be relevant to our predictions. Other than this resources, we modify the evaluation query of AKI from Shirly Wang (2020) to flag on AKI during the first 42 hours.

3 METHODS

3.1 INCLUSION AND EXCLUSION CRITERIA

For simplicity, we focus on predicting adults between the ages 18 to 89 first admission. As our target defined after 48 hours of admission, we will only look on patients with length of stay larger than 48. More than that, we will follow MIMIC-EXTRACT (Shirly Wang (2020)) footsteps, and filter patient with length of stay longer than 240. This is because most patient (70%) stay less than that, and the

[1] Code for this work available here: https://github.com/MarkNoaTAU/MLHC$_Final$

patient that stay longer have heavy tail distribution; they probably represent complex medical cases which we do not aim to cover in our research scope (see Appendix A).

For readmission after 30 days we will also exclude patient that died in-hospital. If a patient died in hospital it is trivial that he will not be readmitted. Although we do not have in hospitality mortality in the moment of predicting the readmission, predicting mortality is easier task that readmission. So, we can use in production two steps prediction algorithm, if wanted. This evaluation was inspired by Kexin Huang (2020).

3.2 Data exploration and preprocessing

Firstly, for preprocessing we used one hot encoder for the categorical variables. We also added indicators variables for overweight and underweight, combine the admission types to emergency admission flag, and extracted indicator for high level of hemoglobin specifically. Other than that, we combined high and weight into BMI. All hourly features were aggregated to point wise statistic's including min, mean, max, std and spread (the difference between the max and min value spotted during the time window).

We computed two steps data imputation. Firstly we impute data based on age-gender group mean value. If all the members in the group have missing values, we used the median of all the cohort for imputation. We did not add indicator values for the missing measurement, as we are already exploring with high dimensional features with small positive labels, and a lot of our features are sparse as it is.

Some features, that seems promising due to their medical meaning, was disappointing during data exploration. For example, DNR stand for do not resuscitate, intuitively it should point on population with high-risk for mortality. Around 5% of hospitalized patient have DNR flag on during their first 42 hours of admission. Out of them 13% will die, not a lot above the general population that stand on 12%. Other features have been found to be much more significant for mortality. For example, the type of admission, the kind of intervention the patient was given, whether or not he experience AKI, and his MELD score (See Appendix B).

Severity scores was important to understand too. First of all, each one of them provide as a base-line as they access the severity of the illness of the patient. Comparing the correlation between the scores, we found out that combining this scores provide additional information (see Figure 1). Indeed this scores rank provide good indicators for mortality, prolong stay and readmission (See Appendix B).

	sirs	sapsii	apsiii	sofa	lods	oasis
sirs	1.00	0.11	0.20	0.20	0.20	0.30
sapsii	0.11	1.00	0.57	0.56	0.70	0.58
apsiii	0.20	0.57	1.00	0.50	0.57	0.46
sofa	0.20	0.56	0.50	1.00	0.64	0.44
lods	0.20	0.70	0.57	0.64	1.00	0.50
oasis	0.30	0.58	0.46	0.44	0.50	1.00

Figure 1: Spreaman correlation matrix of the severity scores over patients. We can see they might provide additional information on top of each other. Mainly SIRS severity score which have low correlation with the other scores

.

As we said before, readmission is the hardest target to predict from the three. As it requires to predict tails events. Finding quality features for readmission was an important challenge. We aimed to have small set of significant features to build small random trees from (to avoid over-fitting). As we said before, each of the severity score was a good feature for readmission. On top of that, we found out that weather or not a patient self paid impact the readmission mean probability. As well as, emergency or urgent admission, DNR and MELD score (more in Appendix B).

3.3 Models

We experiments with Random Forest and Gradient boosting. To address the imbalance of the data we worked with bagging versions of this trees; under-sampling each bootstrap sample to balance

Table 1: Best model performance (on hold-out test)

Target	Balanced Accuracy	F1-score	PR80	Accuracy	AP
Mortality	0.75	0.63	0.46	0.94	0.69
Prolonged stay	0.52	0.71	0.14	0.0	0.42
Hospital Readmission	0.57	0.59	0.1	0.0	0.06

it. Although we wanted, we did not got to experimenting with sophisticated boosting for imbalance data trees, this we had to leave to future work.

For each target, we select the best model, based on its performance on a leave-out test set. Comparing different statistics such as balanced-accuracy, f1-score, PR80 - recall at 80% precision, accuracy and AP; as well as bootstrap confidence evaluation of ROC and PR curves. It was importance to us to look on recall at 80% precision to take into account that we can not tolerance low precision due to alarm fatigue in ICU target prediction. This evaluation was inspired by Kexin Huang (2020).

For each model we tune the hyper-parameters using cross validation on train data, calibration was done on held-out validation data. We also select top K features using highest absolute mean SHAP values on tuned random forests. Where K selected based on the amount of over-fitting on train splits. If we noticed severe over-fitting K was set to 20, else to 75. Than we compared the model performance (after tuning again on the selected features) between the features groups for both RF and GB. The different feature groups compared include: all features extracted; looking only on mean, min, max values of labs, intervention and categorical; and groups of prior-knowledge features (such as the selected homework features, small subset based on the analysis and lists of features extracted from article Nianzong Hou & Wang (2020)).

4 Results

4.1 Mortality

The best model was balanced through under-sampling bagging gradient boosting using basic features from MIMIC-EXTRACT (taking only the mean, min and max of labs and vitals; see table 1 for performance statistics). See appendix D for feature importance results.

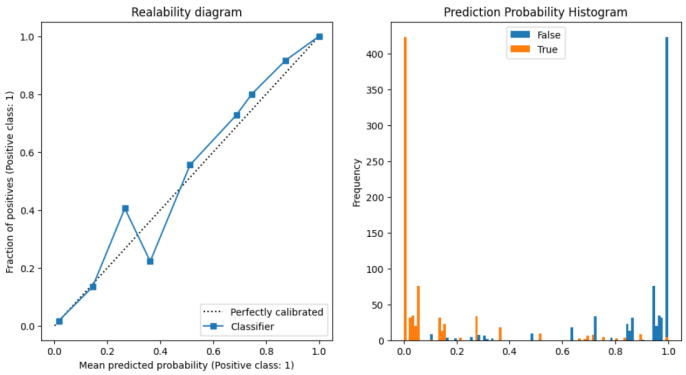


Figure 2: mortality calibration curves

.

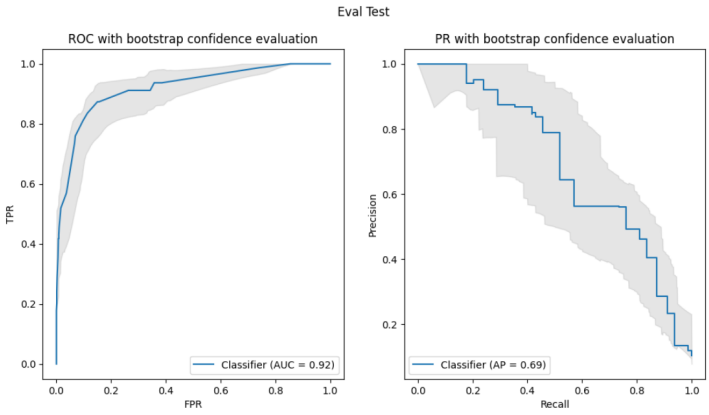


Figure 3: mortality ROC and PR curves

.

4.2 Prolong stay

The best model was under-sampling bagging with random forest using features all extracted features. We can see that on the top importance features using SHAP values (see figure 4) are vaso and vernt intervention alongside vitals and labs statistics. More than that, we can see that variability in the measurements represent by the spread statistic are significant as well.

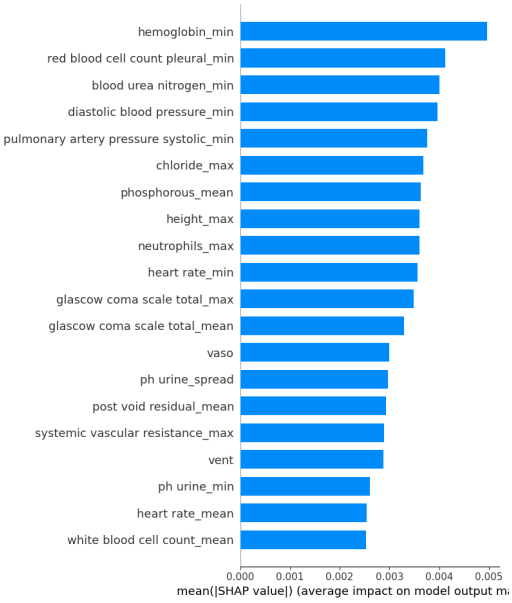


Figure 4: Prolong stay top 20 SHAP values

.

4.3 Hospital readmission

The best model was small gradient boosting model, using depth size 2, and 0.25 l2 regularization to avoid over-fitting. To mitigate over-fitting, we select a small subset of features, that was seems to be significant in the analysis and using prior knowledge. (See appendix D for more detail)

REFERENCES

Lu Shen Li-wei H. Lehman Mengling Feng Mohammad Ghassemi Benjamin Moody Peter Szolovits Leo Anthony Celi Roger G. Mark Alistair E.W. Johnson, Tom J. Pollard. Mimic-iii, a freely accessible critical care database. 2016.

Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The MIMIC Code Repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2017.

Rajesh Ranganath Kexin Huang, Jaan Altosaar. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CHIL 2020 Workshop*, 2020.

Lu He Bing Xie Lin Wang Rumin Zhang Yong Yu Xiaodong Sun Zhengsheng Pan Nianzong Hou, Mingzhe Li and Kai Wang. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. *Journal of Translational Medicine*, 2020.

Geeticka Chauhan Michael C. Hughes Tristan Naumann Marzyeh Ghassemi Shirly Wang, Matthew B. A. McDermott. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. *CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 222–235, 2020.

A APPENDIX

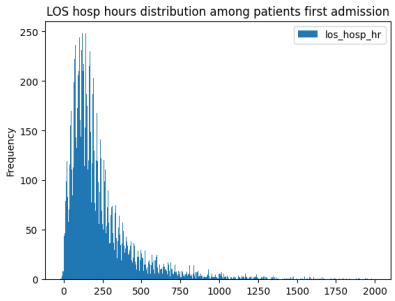
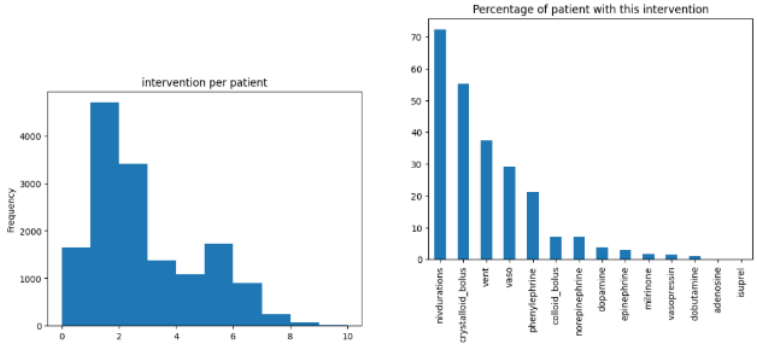


Figure 5: Hourly length of stay; we can see it has heavy tails distribution, which effect our decision to look only on the cohort of patient staid no longer than 420 hours.

.



s

Figure 6: Exploring intervention: on the left number of intervention per patient, on the right, percentage of patient with this intervention.

.

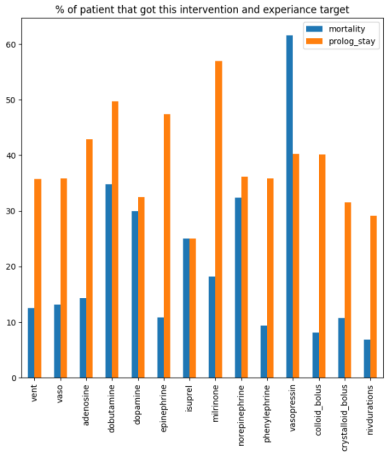


Figure 7: Exploring intervention: % of patient that got this intervention and experienced mortality or prolong stay

.

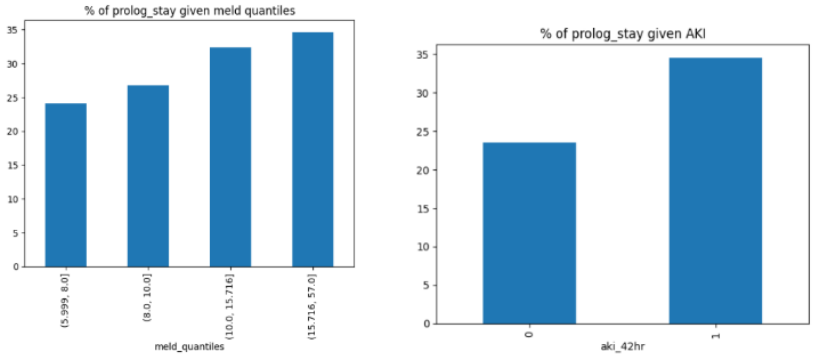


Figure 8: Left, morality given AKI (estimation) in the first 42 hours of admission. Right, mortality given MELD score quantiles.

.

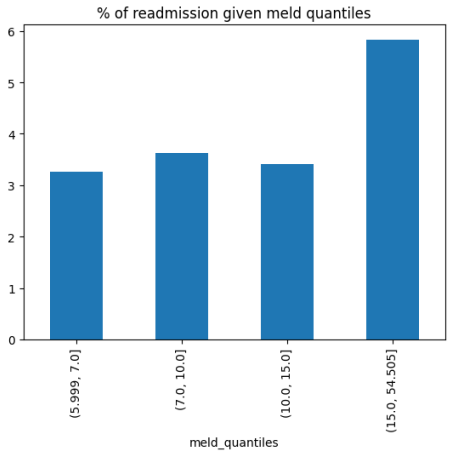


Figure 9: Percentage of readmission given different quantities of MELD score. We can see heaving high MELD score impacts the mean chance of being readmitted, and having low score is below the overall population mean

.

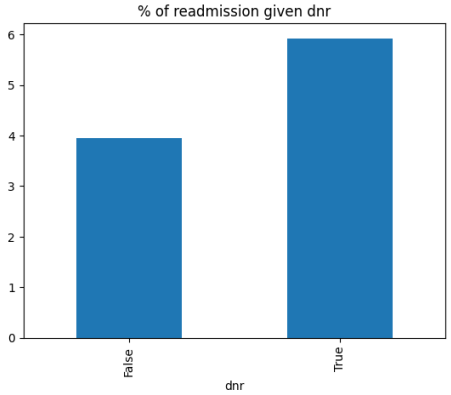


Figure 10: Readmission percentage out of the total patient in the group, given DNR (do not resuscitate order)

.

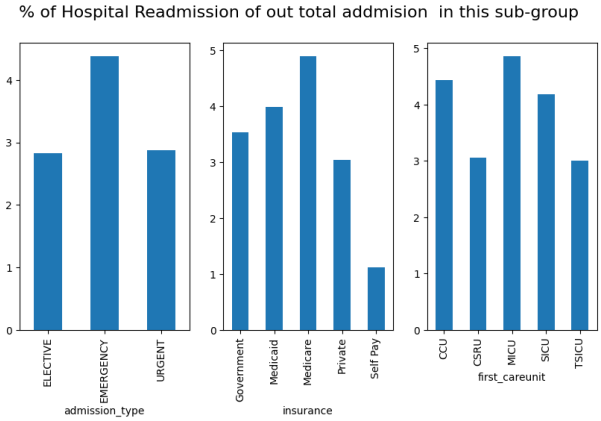


Figure 11: Percentage of readmission out of different population groups. We can see patient that self-paid for admission tend to readmitted less than others. This have an intuitive natural explanation.

.

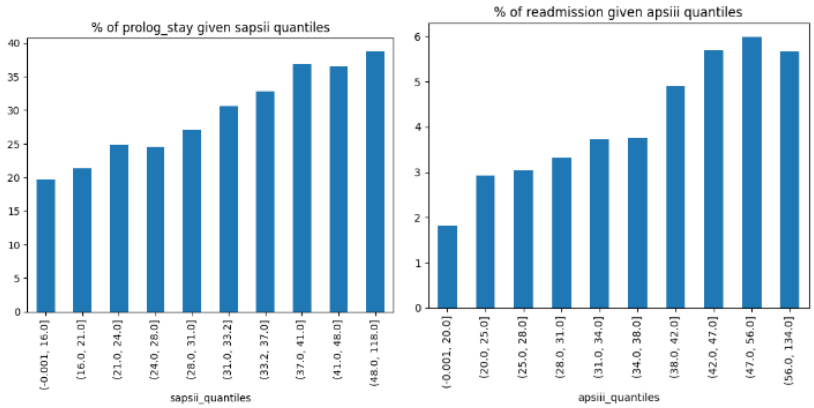


Figure 12: Explore correlation with securities score. On the left prolong stay given different quantilies of SAPS-II score. On the right, Readmission given different quantilies of ASPS-II.

.

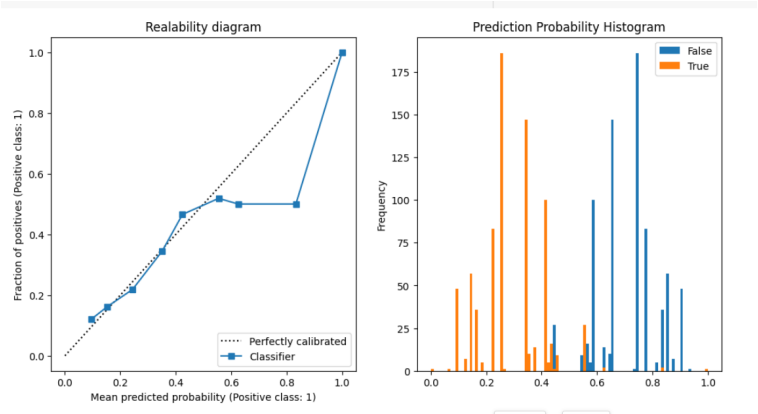


Figure 13: Prolong stay calibration curves

.

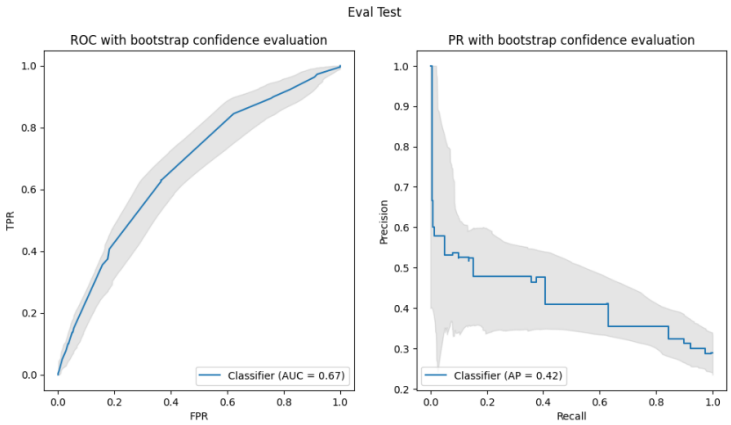


Figure 14: Prolong stay ROC and PR curves

.

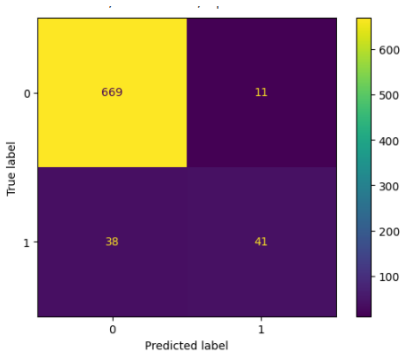


Figure 15: Prolong stay confusion matrix (on test)

.

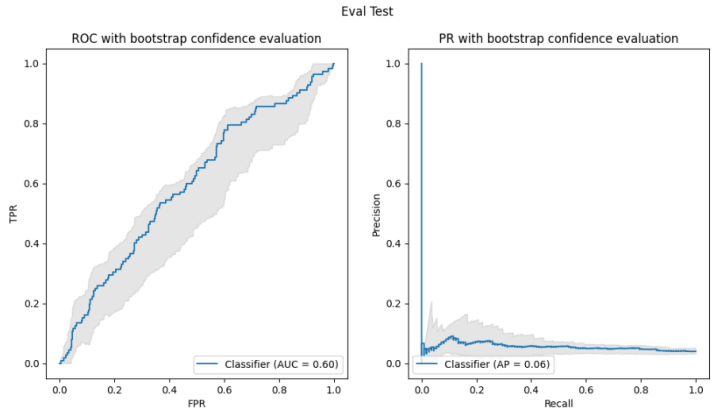


Figure 16: Hospital readmission ROC and PR curves

.

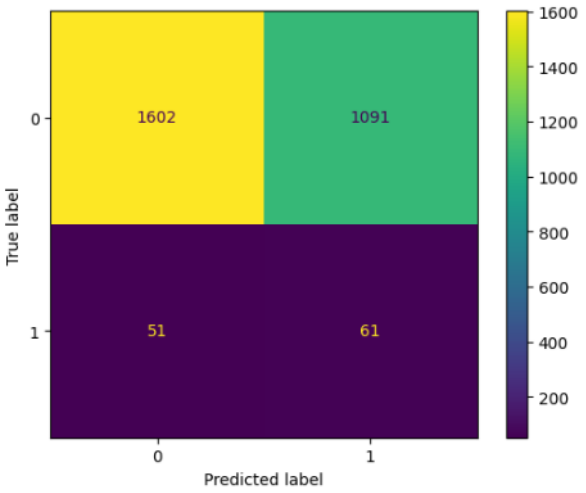


Figure 17: Hospital readmission confusion matrix (on test)

.

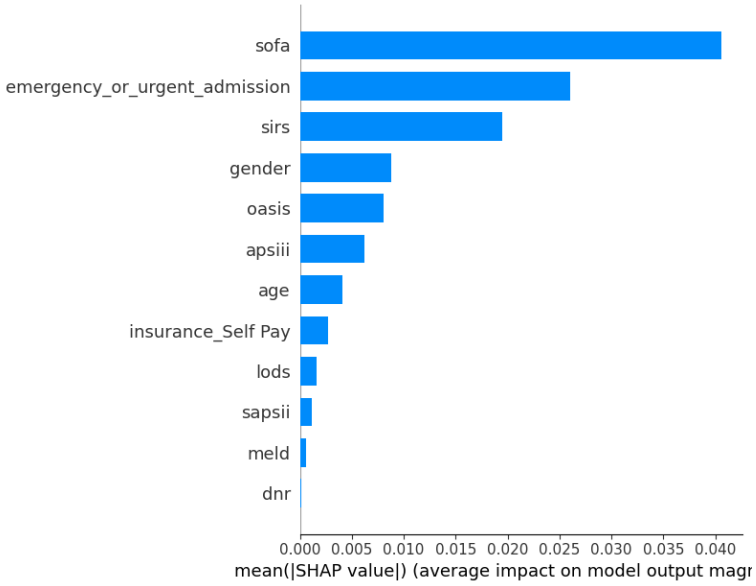


Figure 18: Readmission importance score on the specific prior defined feature set. (Note, SHAP values does not support computation on GB trees, so we looked on this values using Random Forest trees).

.